WEIGHT OF EVIDENCE (WOE) and INFORMATION VALUE (IV)

- Logistic regression model is one of the most commonly used statistical technique for solving binary classification problem.
- The two concept - weight of evidence (WOE) and information value (IV) evolved from same logistic regression technique.

Weight of Evidence (WOE) -

→ WOE tells the predictive power of an independent variable in relation to the dependent variable. Since it is evolved from credit scoring world, it is generally described as a measure of the seperation of good customer and bad customers. Good customers refers to the customers who paid loan back and Bad customers who defaulted on loan.

$$\boxed{WOE = \ln\left[\frac{\text{Distribution of Good}}{\text{Distribution of Bad}}\right]}$$

ln → Natural log.
Distribution of Good → % Good customers in particular group.
Distribution of Bad → % Bad customers in particular group.

→ Positive WOE → Distribution of Good > Distribution of bad.
Negative WOE → Distribution of Good < Distribution of bad

Hint → log of a number > 1 means positive value.
log of a number < 1 means negative value.

In general terms, $\boxed{WOE = \ln\left(\frac{\%\text{ of non-events}}{\%\text{ of events}}\right)}$

Steps to Calculate WOE → ① For continuous variable, split data into bins.
② Calculate the number of event & non event in each bins
③ Calculate the % of events & % non event in each bins
④ Calculate WOE by taking log of division (% non event & % event)

NOTE → For Categorical variable, we do not need to split the data (Ignore step 1. & follow remaining)

Example—

Range	Bins	Non Event	% Non Event	Event	% Event	WOE	IV
0-50	1	197	5.4%	20	5.9%	-0.0952	0.0005
51-100	2	450	12.3%	34	10.1%	0.2002	0.0045
101-150	3	492	13.4%	39	11.5%	0.1522	0.0029
151-200	4	597	16.3%	51	15.1%	0.0774	0.0009
204-250	5	609	16.6%	54	16.0%	0.0401	0.0003
250-300	6	582	15.9%	55	16.3%	-0.0236	0.0001
301-350	7	386	10.5%	41	12.1%	-0.1405	0.0022
351-400	8	165	4.5%	23	6.8%	-0.4123	0.0095
>401	9	184	5.0%	21	6.2%	-0.2123	0.0025
	Tota	3662	[illegible]	338			0.0234

Rules related to WOE -

1. Each bins should have atleast 5% of observations.
2. Each bins should be non-zero for both non-events and events.
3. WOE should be distinct for each category. Similar group should be aggregated
4. WOE should be monotonic, i.e, either growing or decreasing with groupings.
5. Missing value are binned seperately.

How to check correct binning with WOE -

1. The WOE should be monotonic i.e, either growing or decreasing with bins. We can plot WOE and check linearity on graph.
2. Perform WOE transformation & check with logistic regression output.

Terminology related to WOE -

1. Fine classing - Applied to all continuous variables and those discrete variable with high cardinality. This is the process of initial binning into typically between 20 and 50 fine granular bins.

   To summarize create 10/20 bins for a continuous independent variable & calculate WOE and IV of a variable

2. Coarse classing - Combine adjacent categories with similar WOE scores.

Usage in Model -

1. Continuous Independent variable - First create bins for that variable and then combine categories with similar WOE values and replace categories with WOE values. Use WOE values rather than input values.

   Eg -
   if age => 10 then WOE-age = -0.03012
   if age >= 20 then WOE-age = -0.07689
   if age == NULL then WOE-age = 0.34616

2. Categorical independent variable - Combine categories with similar WOE and then create new categories of an independent variable with continuous WOE values. Use WOE values rather than raw categories in model. Transformed variable will be continuous variable with WOE value. It is same as any continuous value.

Why combine categories with similar WOE?

→ It is because the categories with similar WOE have almost same proportion of events and non-events. In other word, the behavior of both the categories is same.

Information Value (IV) -

- IV is one of the most useful technique to select important variable in a predictive model.
- It help us to rank variables on the basis of importance.

→ $$\boxed{\text{Information Value} = \sum \left(\frac{\%\text{ of non-events}}{\%\text{ of events}} \right) * \text{WOE}}$$

INFORMATION VALUE	VARIABLE PREDICTIVENESS
Less than 0.02	Not useful for prediction (Not useful for modelling)
0.02 to 0.1	Weak predictive power (weak relation to Good Bad)
0.1 to 0.3	Medium predictive power (medium strength relation between good & bad)
0.3 to 0.5	strong predictive power (strong relation betw" good bad)
> 0.5	Suspicious predictive power (check once again)

→ IV increases as Bins / groups increases for an independent variable. Be careful when there are 20 bins as some bins may have very few number of events and non-events.

→ IV is not an optimal feature (variable) selection method when we are building a classification model other than binary logistic regression as conditional log odds is highly related to calculation of weight of evidence.

→ Random forest can detect non-linear relationship very well so selecting variable via Information Value and using them in random forest might not produce most accurate and robust predictive model.

Advantage of WOE & IV

① Main practical use of WOE is for encoding, where we can replace the classes with their associated value. For example, suppose in a dataset we found we can replace "Male" with 0.98383 and "Female" with -1.582.

② Another positive outcome of using WOE is to reduce the number of columns of the input used for training a model. Imagine we have a categorical variable with 10 different classes and we performed one-hot encoding, we will end up with 10 columns with mostly 0 as values. Using WOE, classes are replaced by their associated WOE values.

③ As for IV, it provide relationship between independent & dependent variables. With help of WOE & IV we can engineer meaningful features.

If target variable is continuous, WOE and IV?

→ we can find WOE and IV but we need to modify the formula.

$$\text{Modified WOE} = \ln\left(\frac{\%\text{ of } Y}{\%\text{ observation}}\right)$$

$$\text{Modified IV} = \Sigma\left(\left(\%\text{ of } Y - \%\text{ Observation}\right) * \text{Modified WOE}\right)$$

Steps - 1) Split Continuous independent variable (X) into 10 or 20 buckets (called variable 'rank'). If we have categorical independent variable, we don't need to split as they are already categorized.

2) Calculate min and max of X by rank. Compute sum of target variable (Y) by rank. Let's name it as 'Sum Y'.

3) Calculate total count & % of observation falling in each bucket of rank variable

4) Calculate % Y which is calculated by Sum Y / Σ Sum Y.

5) WOE = ln(%Y / % Obs). % Obs represent percentage of observation (step3)

6) IV = Σ((%Y - % Obs) * WOE)

Bins	Lower limit (Min X)	Upper limit (Max X)	Sum(Y)	N	% observation	% Y = (SumY/ΣSumY)	WOE = ln(%Y / %Obs)	IV = (%Y - %Obs) * WOE
1	0.00	0.21	21.32	252	10%	3.7%	-1.00	0.06
2	0.21	0.35	25.52	254	10%	4.4%	-0.82	0.05
3	0.35	0.44	31.64	252	10%	5.4%	-0.61	0.03
4	0.44	0.56	32.04	252	10%	5.5%	-0.51	0.03
5	0.56	0.66	32.44	254	10%	5.6%	-0.58	0.03
6	0.66	0.76	45.44	252	10%	7.8%	-0.24	0.01
7	0.76	0.87	61.30	254	10%	10.6%	0.06	0.00
8	0.87	0.97	86.42	253	10%	14.9%	0.40	0.02
9	0.97	1.14	109.52	253	10%	18.9%	0.63	0.06
10	1.14	1.98	134.93	254	10%	23.2%	0.85	0.11
								0.38

Work flow of WOE →